(Rascunho) O efeito do envelhecimento populacional brasileiro no recrutamento militar: uma análise orientada a dados com Python

Lázaro Damasceno

18 de janeiro de 2026

Resumo

Resumo em PT-BR.

Resumo

Resumo em EN.

1 Introdução

2 Referencial teórico

2.1 Política federal de dados abertos

2.2 Serviço militar

2.3 Envelhecimento da população

3 Metodologia

Esta seção é divida em duas subseções. A subseção 3.1 detalhará como a linguagem de programação Python foi usada. A subseção 3.2 detalhará como foi o tratamento dos dados do EB e do IBGE.

3.1 Python para ciência dos dados e aprendizado de máquina

Python foi escolhida como a linguagem de programação a ser usada para analisar os dados do serviço militar do EB e das projeções populacionais do IBGE.

As bibliotecas de Python usadas foram polars, matplotlib e scikit-learn. A escolha do polars ao invés do pandas foi motivada pela sua API moderna construída com Rust do polars. Essa tem um poder de processamento de dados mais potente do que o tradicional pandas.

Outro ponto forte do polars é sua abordagem funcional associada a seu lazy evaluation. O conjunto é ideal, pois torna possível executar comandas diferentes em uma estrutura encadeada sem os problemas de lentidão associados ao eager evaluation para conjuntos de dados grandes.

O matplotlib foi usado para a geração e customização das figuras. A biblioteca define o tamanho das figuras pela largura e altura, respectivamente, em polegadas. Um polegada equivale a 2,54 cm. O tamanho padrão das figuras é 10x5 polegadas. Caso seja necessário usar outro tamanho, ele será adotado.

O scikit-learn será usada na adaptação dos dados para o aprendizado de máquina e a posteriori modelagem baseada no modelo escolhido por ser o mais adequado.

3.2 Conjuntos de dados

O conjunto de dados do EB consta na URL https://dados.gov.br/dados/conjuntos-dados/servico-militar. Até 17/01/2026, são 18 arquivos em CSV. O ínicio da série histórica começa em 2007 e termina em 2024. O arquivo em CSV com os dados do serviço militar de 2025 ainda não foram adicionados pelo EB.

Dos dados existentes, eles totalizam 4,9 GB. Para que pudessem ser usados com polars, foram transferidos para a plataforma Databricks, que lida com Big Data. Nela, foram transformados em arquivos do tipo Parquet, cujo tamanho total é 165,1 MB. A redução foi de 96,7%.

Todas as colunas do conjunto de dados constam na documentação do recurso. O documento consta na URL https://dadosabertos.eb.mil.br/arquivos/sermil/dicionario-dados-sermil.html.

Das 21 colunas originais, 6 foram removidas, ou seja, uma redução 28,57%. Elas são as colunas relativas ao tamanho da cabeça, largura da cintura, tamanho do calçado, religião dos alistados, concomitantemente, à Junta de serviço Militar e o município e a unidade federativa dela.

Com os arquivos em Parquet disponíveis, a análise dos dados do serviço militar foi dividida em três etapas: ETL e data wrangling, análise exploratória de dados e aprendizado de máquina. Polars foi usado nas duas primeiras etapas e scikit-learn na última.

Os códigos de todas as análises constam na URL https://github.com/LazaroDamasceno/Rascunho-Artigo/tree/main/analise_dados. Eles foram gerados em um ambiente virtual do Anaconda na sua versão reduzida, o miniconda.

A análise dos dados do serviço militar inicia-se com a etapa de ETL e data

wrangling. O primeiro passo foi descobrir quais colunas valores nulos. As colunas PESO, ALTURA e DISPENSA têm valores nulos, respectivamente, 18.598.039, 18.588.096 e 474.630. Originalmente, o dataframe tem 26.569.408 de linhas. Ou seja, representam, respectivamente, 69,99%, 69,96% e 1,78% da quantidade de linhas originais.

A tabela 1 detalha por ano a quantidade de valores nulos das colunas PESO, ALTURA e DISPENSA.

Tabela 1: Quantidade de valores nulos por ano e categoria

ANO DO ALISTAMENTO	PESO	ALTURA	DISPENSA
2007	81.15	80.7	0.0
2008	81.6	81.47	0.0
2009	78.75	78.74	0.0
2010	79.92	79.92	0.0
2011	79.71	79.71	0.0
2012	81.24	81.24	0.0
2013	80.15	80.15	0.0
2014	80.2	80.19	0.0
2015	79.52	79.52	0.0
2016	79.34	79.35	0.0
2017	74.84	74.82	32.7
2018	68.11	68.11	0.0
2019	0.0	0.0	0.0
2020	68.1	68.12	0.0
2021	67.53	67.53	0.0
2022	70.52	70.52	0.0
2023	56.78	56.78	0.0
2024	0.0	0.0	0.0

Como exposto pela tabela 1, os valores nulos por ano das colunas PESO e ALTURA representam massivamente a quantidade total dos dados anuais. Neste sentido, foram removidas.

Sobre a coluna remanescente DISPENSA, apenas o ano de 2017 tem valores nulos. Eles representam quase $\frac{1}{3}$. Foi descoberto que DISPENSA tem como valores "Com dispensa", "Sem dispensa", "null"e null. O valor textual "null"é o valor nulo embalado no tipo String.

Para corrigir o erro lógico, todos os valores "null"foram transformados em valores nulos em todas as colunas, pois poderia estar presente em outras colunas.

Assim, foi descoberto que após a mudança, apenas a coluna DISPENSA tinha

valores nulos. Como essa possuí uma lógica binária de ter sido dispensado ou não, os valores nulos foram removidos. Com a remoção, a quantidade de linhas é 26.094.773.

Devido à logica binário da coluna DISPENSA, seus dados foram transferidos para uma nova coluna: convocacao. Ele inverte a lógica da primeira, focando na convocação ou não para o serviço militar. Após isto, DISPENSA foi removida.

Após isso, as colunas ANO_NASCIMENTO e VINCULACAO_ANO foram transformadas de String para Int16, concomitantemente, UF_NASCIMENTO e UF_RESIDENCIA foram transformadas para o tipo categórico. Outras duas colunas foram removidas: MUN_NASCIMENTO e MUN_RESIDENCIA.

As colunas PAIS_NASCIMENTO e PAIS_RESIDENCIA foram usadas para verificar a quantidade respectiva de alistados que nasceram e residiam no Brasil no momento do alistamento.

A quantidade de alistados que moravam no Brasil no momento da seleção militar era de 97.06%, enquanto os outros representavam 2,94%. No tocante à residência, 95,58% residiam no Brasil; os outros, 4,42%.

Considerando a baixa representatividade dos nascidos e residentes no exterior, apenas os nascidos e residentes em solo nacional foram selecionados. Assim, pode-se remover as colunas PAIS_NASCIMENTO e PAIS_RESIDENCIA. A quantidade de linhas foi reduzida para 25.935.637.

A próxima coluna a ser tratada é ESTADO_CIVIL. Ela contém os seguintes valores: "Solteiro", "Casado", "ViÃºvo", "Viúvo", "Separado Judicialmente", "Outros", "Desquitado", "Divorciado".

Adicionalmente, a inconsistência "ViÃºvo"foi renomeada para "Viúvo". Os valores "Separado Judicialmente"e "Desquitado"foram renomeados como "Divorciado", pois legalmente são divorciados.

A tabela 2 mostra a quantidade percentual dos status civis dos alistados.

Tabela 2: Quantidade dos status civis dos alistados

ESTADO CIVIL	**QUANTIDADE (%)**
Solteiro	97,55
Casado	1,4
Divorciado	0,18
Viúvo	0,01

Conforme mostra a tabela 2, a quantidade de alistados que são solteiro é a quase a totalidade, de modo que os demais estados civis representam apenas 2,45% do total. Como consequência, apenas os alistados solteiros foram selecionados e coluna ESTADO_CIVIL foi removida. A quantidade de linhas foi reduzida para 25.298.928.

A próxima coluna a ser tratada é a SEXO. Ela contém apenas "M"para os homens e "F"para as mulheres. Como 99,95% dos alistados são homens e as mulheres representam o valor insignificante de 0,05%, apenas os homens foram escolhidos. Assim, a coluna SEXO foi removido. A quantidade de linhas foi reduzida para 25.287.523.

Para a coluna ZONA_RESIDENCIAL, que contém apenas os valores "Rural"e "Urbano", descobriu-se que, no geral, 87,89% dos alistados residem na zona urbana, enquanto 12,11% vivem na zona rural. Como não havia necessidade de remoção, ZONA_RESIDENCIAL foi transformada no tipo categórico.

4 Discussão

4.1 Análise univariada

4.2 Análise bivariada

4.3 Análise multivariada

5 Resultados

6 Considerações finais